Bioinformatic approaches to regulatory genomics and epigenomics

376-1347-00L | week 05

Pierre-Luc Germain

Plan

- Debriefing on the assignment

- The ‘histone code’ & functional elements

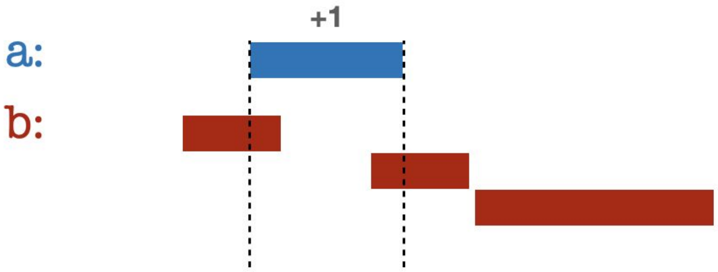
- More on overlaps and comparing signals

Debriefing on the assignments

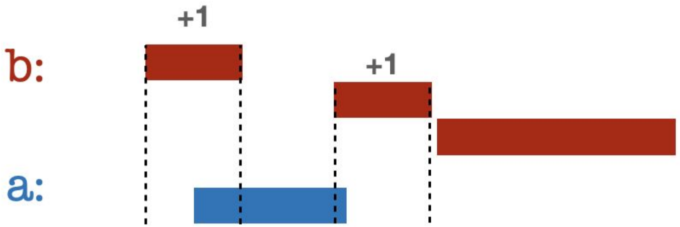
Symmetry of overlaps:

a: query
b: subject

overlapsAny(a, b)
findoverlaps(a, b)



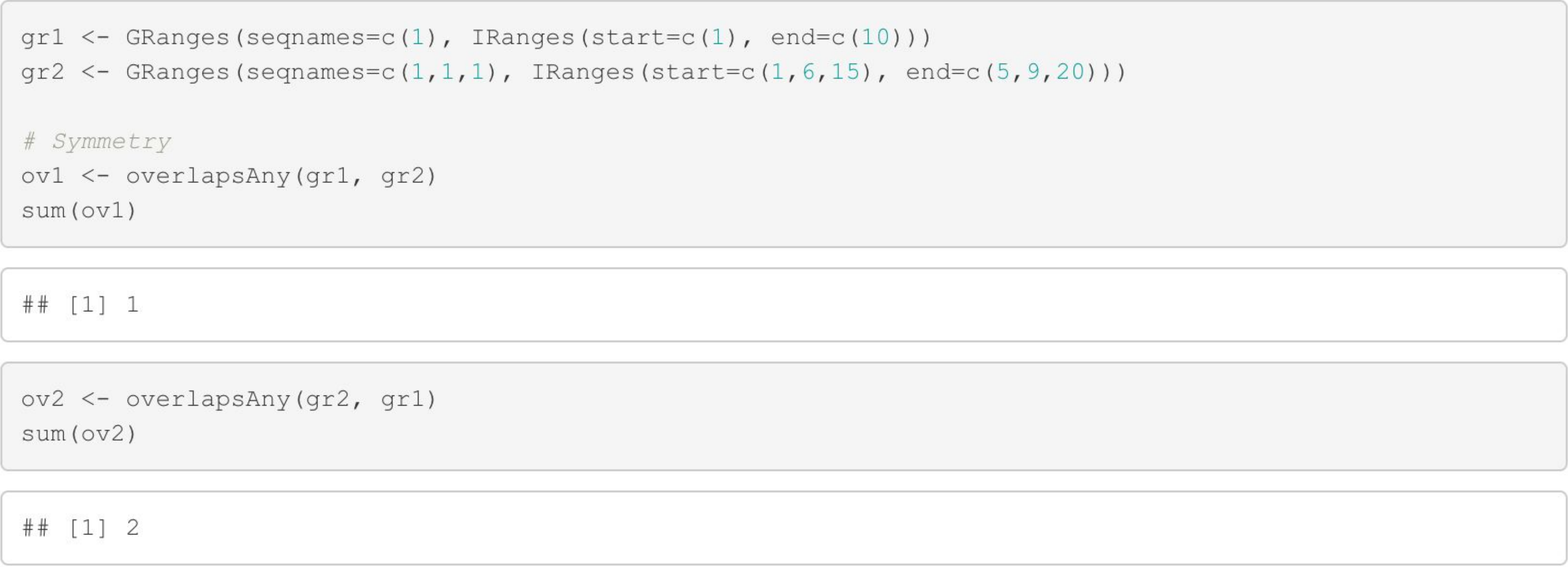
sum(overlapsAny(a, b))=1



sum(overlapsAny(b, a))=2

Debriefing on the assignments

Symmetry of overlaps:



```r
gr1 <- GRanges(seqnames=c(1), IRanges(start=c(1), end=c(10)))
gr2 <- GRanges(seqnames=c(1,1,1), IRanges(start=c(1,6,15), end=c(5,9,20)))

# Symmetry
ov1 <- overlapsAny(gr1, gr2)
sum(ov1)
```

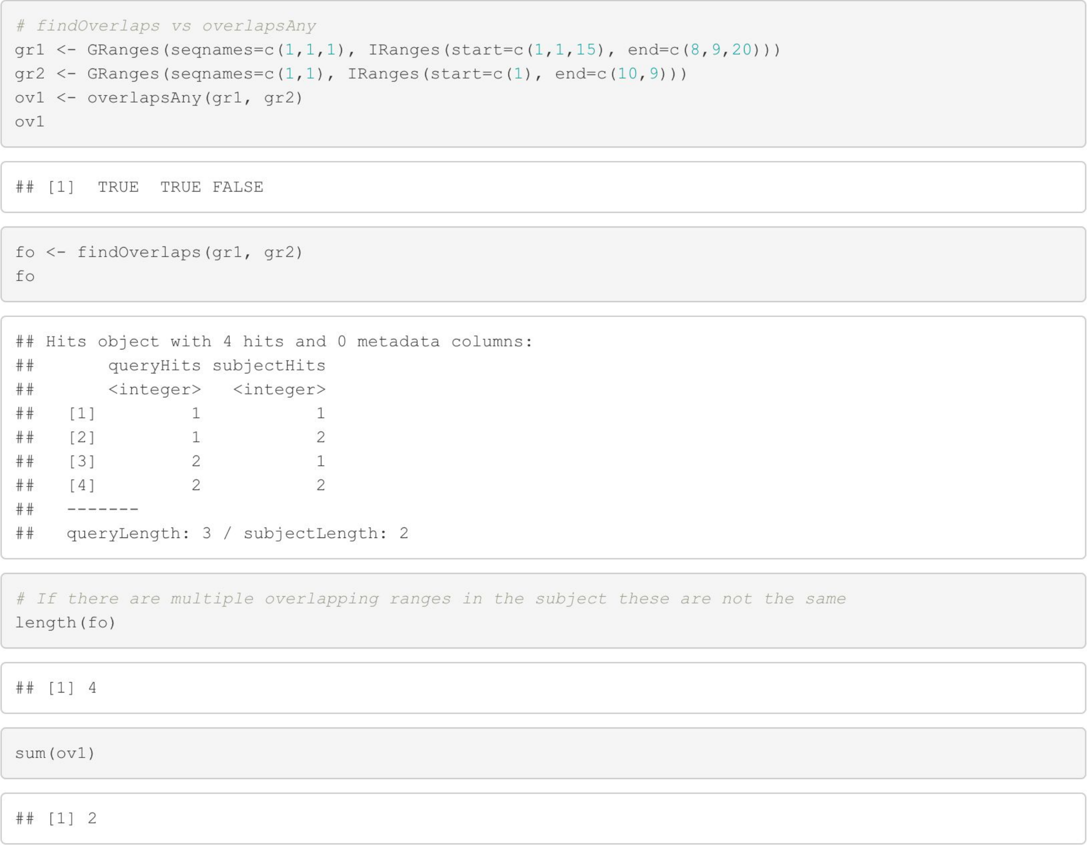
```
## [1] 1
```

```r
ov2 <- overlapsAny(gr2, gr1)
sum(ov2)
```

```
## [1] 2
```

Debriefing on the assignments

Difference `findOverlaps()` vs `overlapsAny()`:



```
# findOverlaps vs overlapsAny
gr1 <- GRanges(seqnames=c(1,1,1), IRanges(start=c(1,1,15), end=c(8,9,20)))
gr2 <- GRanges(seqnames=c(1,1), IRanges(start=c(1), end=c(10,9)))
ov1 <- overlapsAny(gr1, gr2)
ov1
```

```
## [1]  TRUE  TRUE FALSE
```

```
fo <- findOverlaps(gr1, gr2)
fo
```

```
## Hits object with 4 hits and 0 metadata columns:
##       queryHits subjectHits
##       <integer>   <integer>
##   [1]         1           1
##   [2]         1           2
##   [3]         2           1
##   [4]         2           2
##   -------
##   queryLength: 3 / subjectLength: 2
```

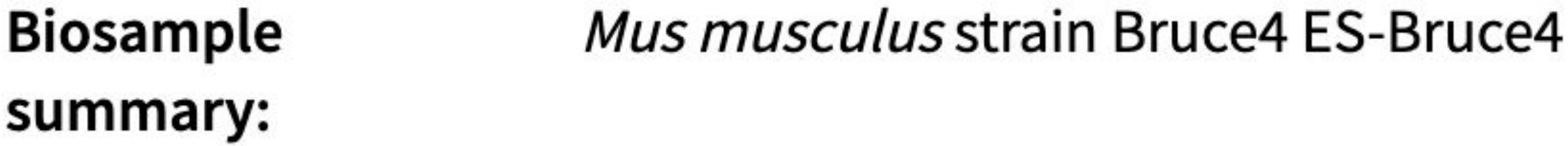
```
# If there are multiple overlapping ranges in the subject these are not the same
length(fo)
```

```
## [1] 4
```

```
sum(ov1)
```

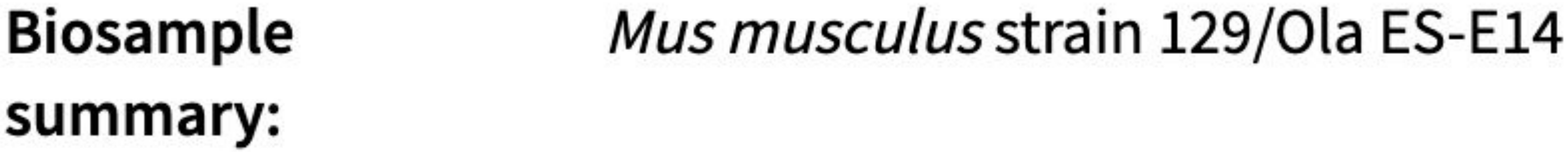
```
## [1] 2
```

p300:

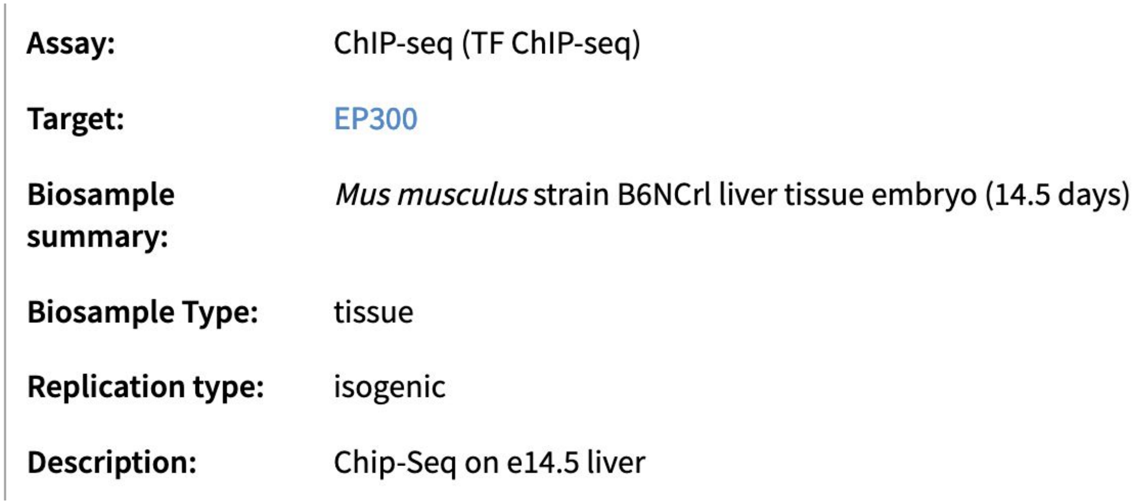


Biosample summary: *Mus musculus* strain Bruce4 ES-Bruce4

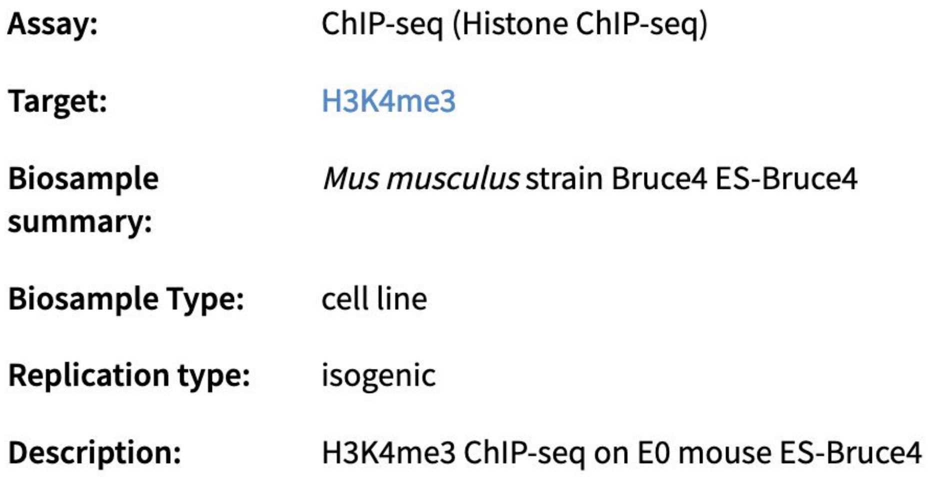
H3K4me1:



Biosample summary: *Mus musculus* strain 129/Ola ES-E14

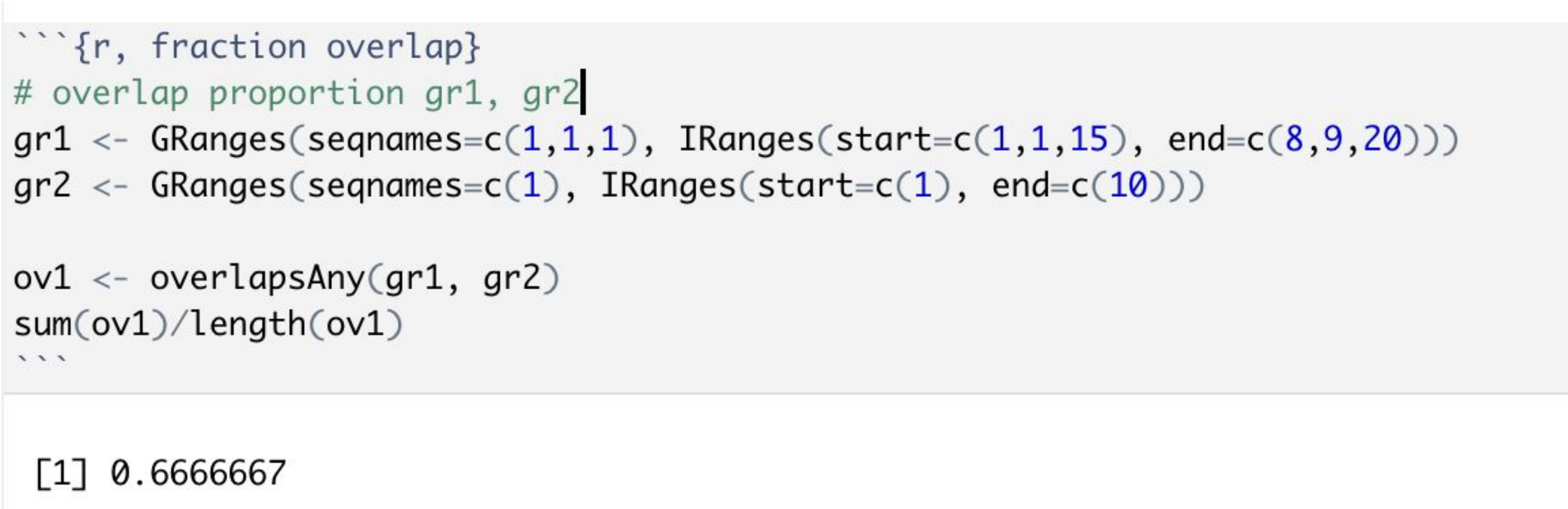


Assay:	ChIP-seq (TF ChIP-seq)
Target:	EP300
Biosample summary:	*Mus musculus* strain B6NCrl liver tissue embryo (14.5 days)
Biosample Type:	tissue
Replication type:	isogenic
Description:	Chip-Seq on e14.5 liver



Assay:	ChIP-seq (Histone ChIP-seq)
Target:	H3K4me3
Biosample summary:	*Mus musculus* strain Bruce4 ES-Bruce4
Biosample Type:	cell line
Replication type:	isogenic
Description:	H3K4me3 ChIP-seq on E0 mouse ES-Bruce4

Overlapping proportion:

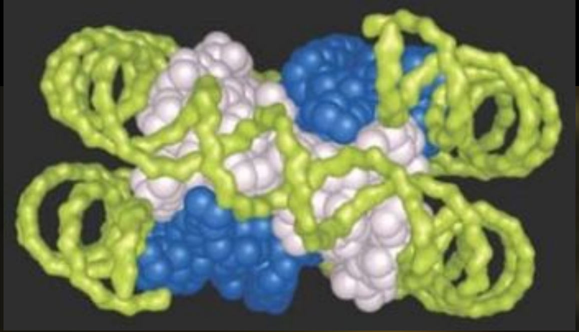


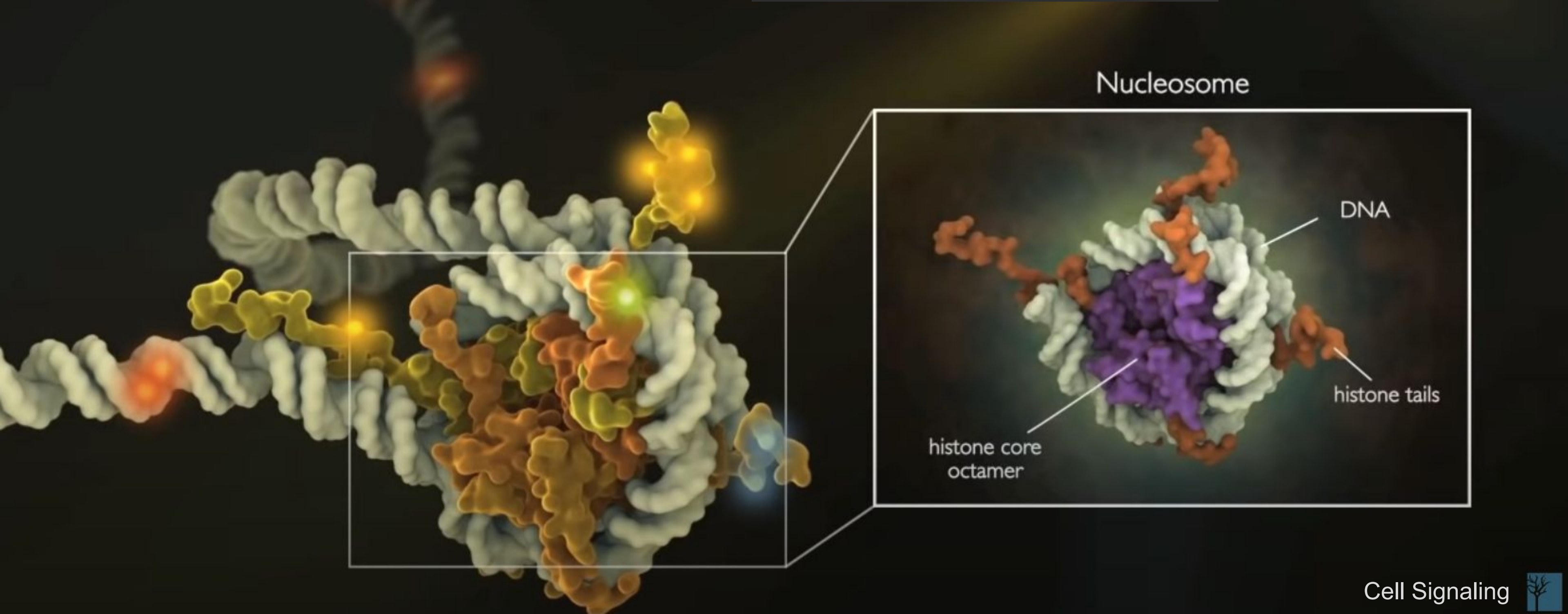
~~~
```{r, fraction overlap}
# overlap proportion gr1, gr2
gr1 <- GRanges(seqnames=c(1,1,1), IRanges(start=c(1,1,15), end=c(8,9,20)))
gr2 <- GRanges(seqnames=c(1), IRanges(start=c(1), end=c(10)))

ov1 <- overlapsAny(gr1, gr2)
sum(ov1)/length(ov1)
```
~~~

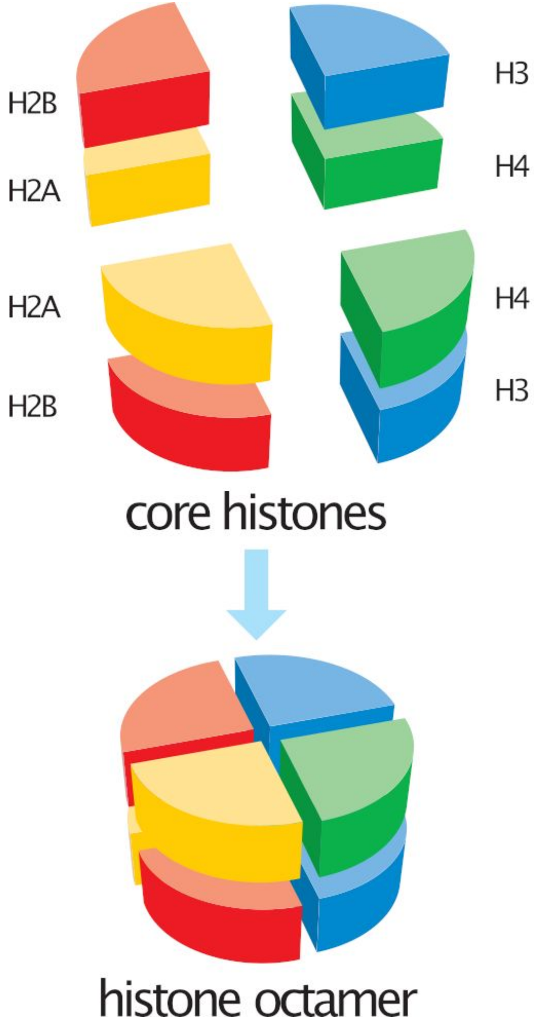
```
[1] 0.6666667
```

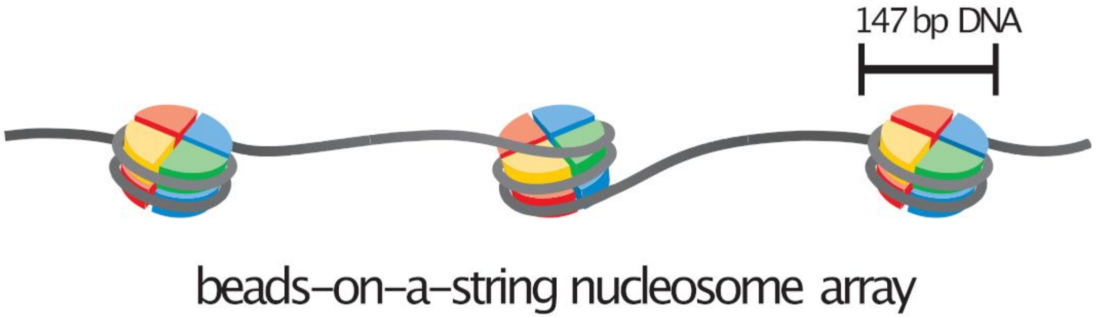
Nucleosome

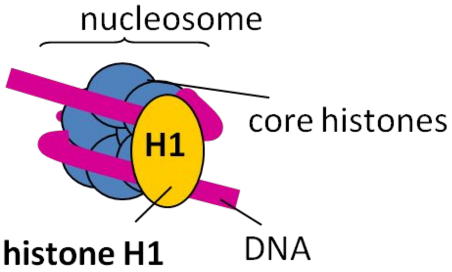
(Krebs et al., 2018)





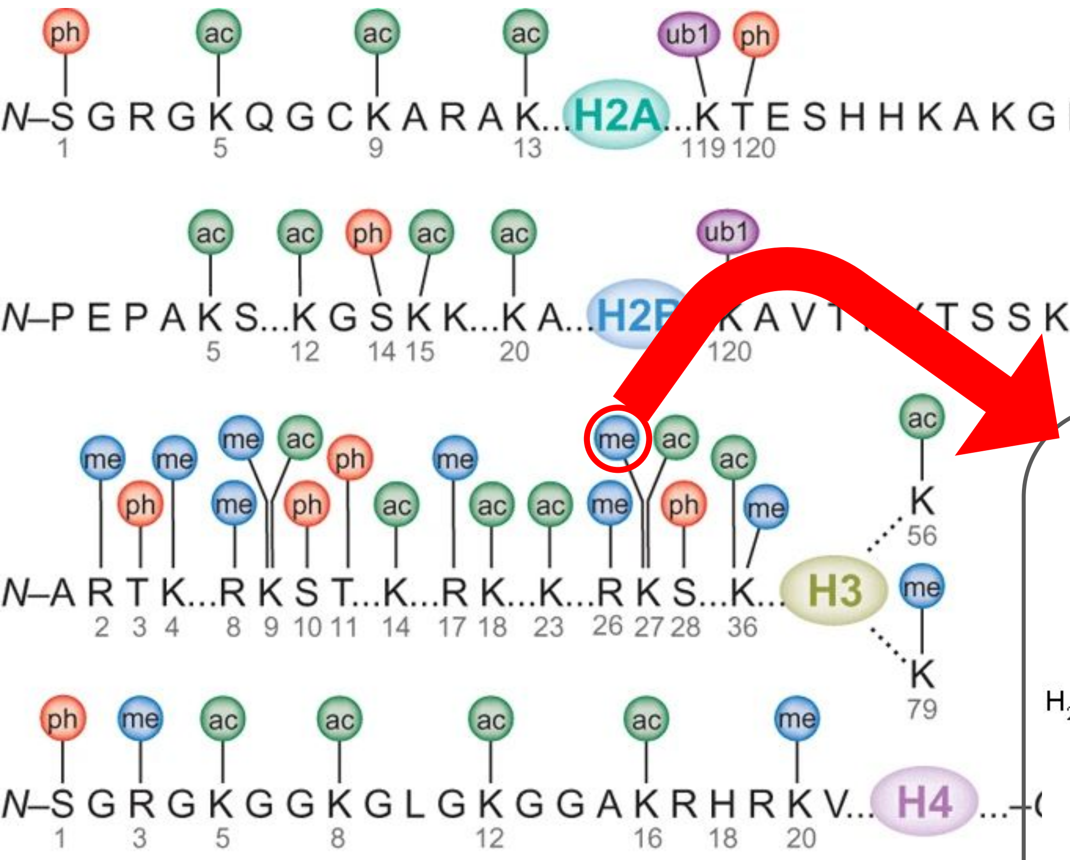






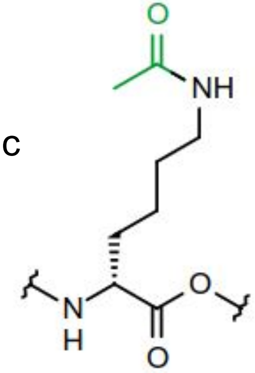
(Adapted from David O Morgan - The Cell Cycle. Principles of Control. Wikimedia Commons)

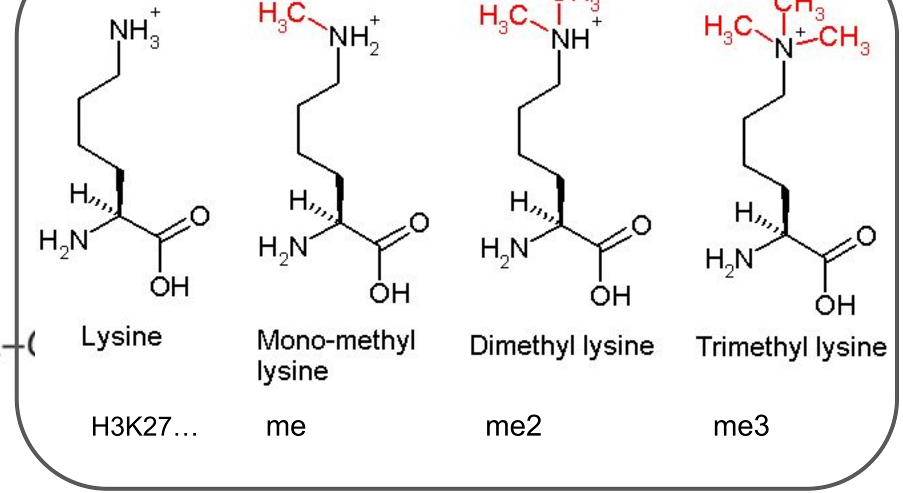
Many residues on the histone tails can be post-translationally modified


ph ac ac ac ub1 ph
N–S G R G K Q G C K A R A K...H2A...K T E S H H K A K G K...–C
1 5 9 13 119 120
ac ac ph ac ac ub1
N–P E P A K S...K G S K K...K A...H2B...K A V T K T S S K...–C
5 12 14 15 20 120
me ph me me me ac ph me ac ac ac me ac me ph ac me ac me
N–A R T K...R K S T...K...R K...K...R K S...K...H3
2 3 4 8 9 10 11 14 17 18 23 26 27 28 36
K 56
K 79
ph me ac ac ac ac me
N–S G R G K G G K G L G K G G A K R H R K V...H4...–C
1 3 5 8 12 16 18 20

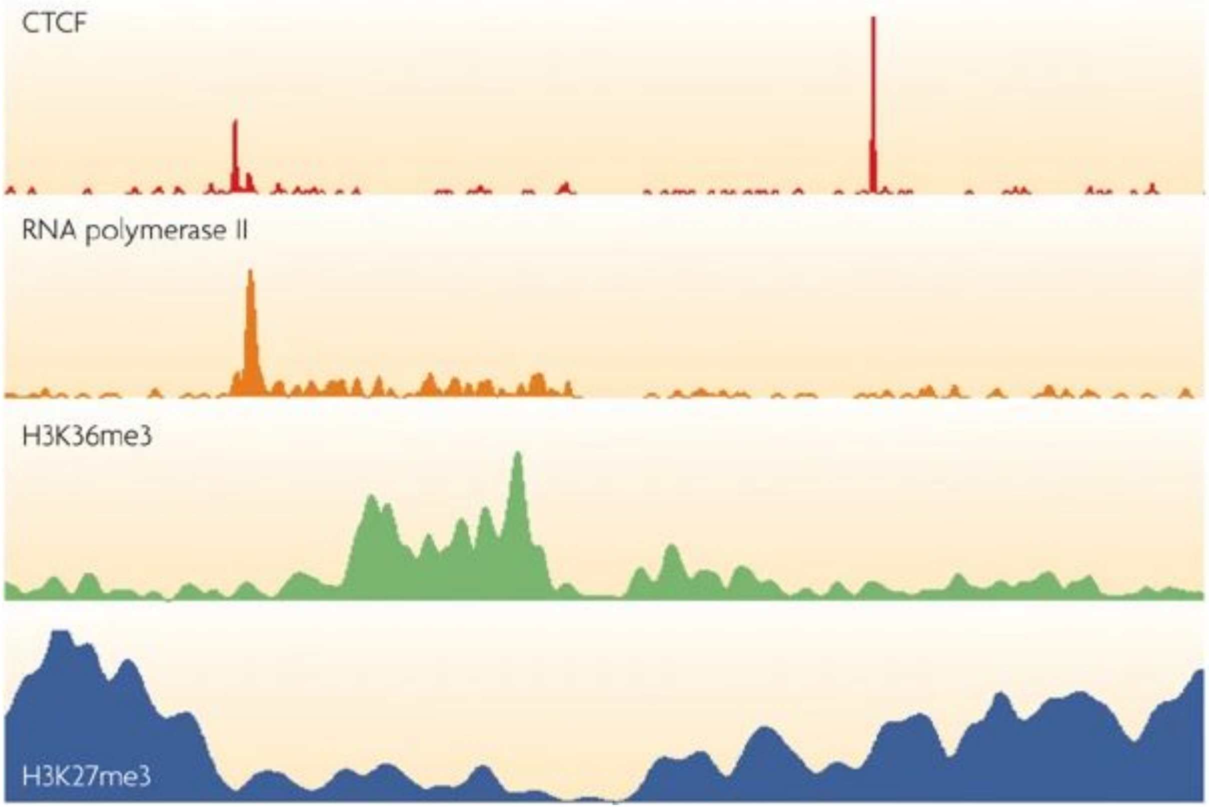
(Bhaumik, Smith and Shilatifard 2007)

H3K27ac





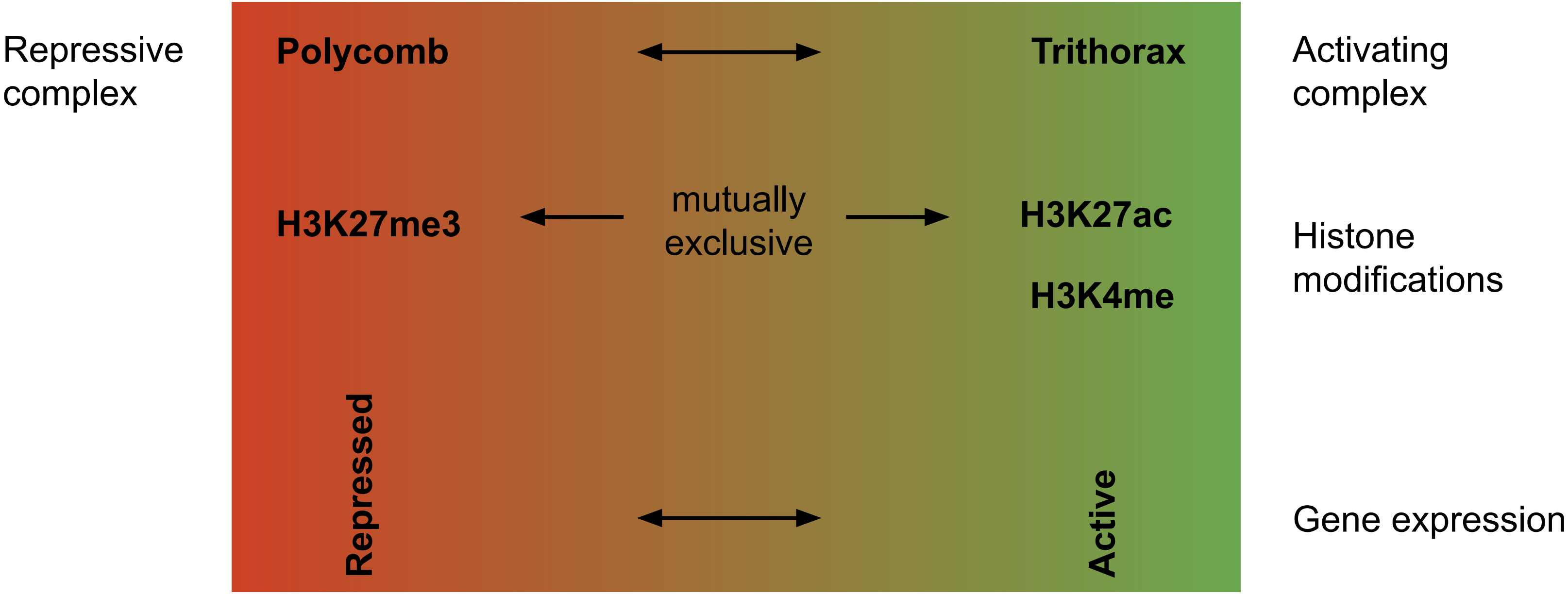
Some histone modifications appear to be very localized, e.g. happening on a specific nucleosome, while most are much more broadly distributed



The strategy of calling 'peaks' must therefore be adapted
(e.g. "broad" option of most peak-callers)

(Park, Nat Rev Gen 2009)

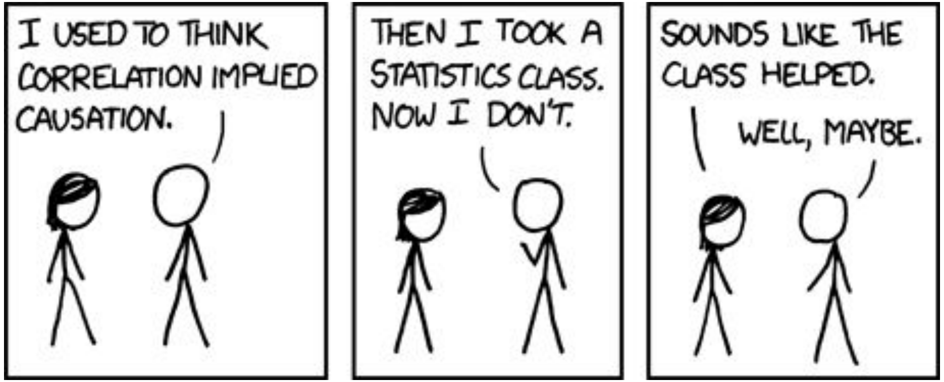
There is a very strong association of certain histone marks and activation or repression



But which comes first?

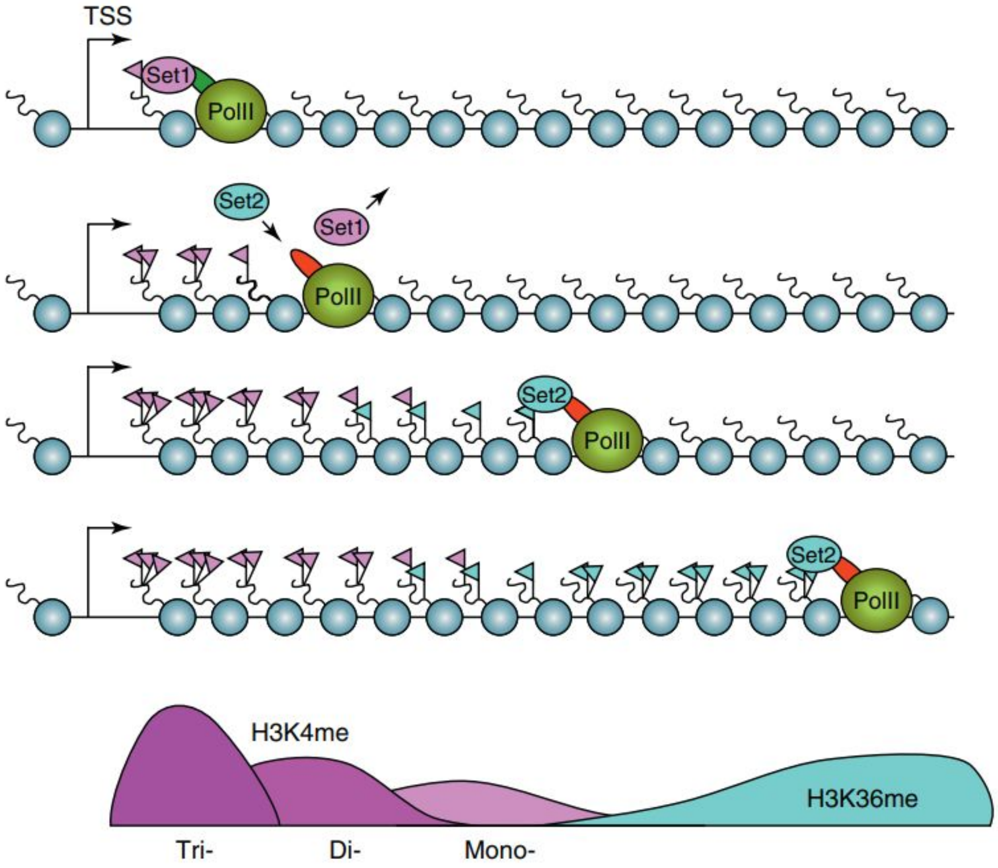
Causality or correlation?

Are histone modifications **responsible** for activation/repression, or are they merely associated **side-effects**?


I USED TO THINK CORRELATION IMPLIED CAUSATION.
THEN I TOOK A STATISTICS CLASS. NOW I DON'T.
SOUNDS LIKE THE CLASS HELPED.
WELL, MAYBE.

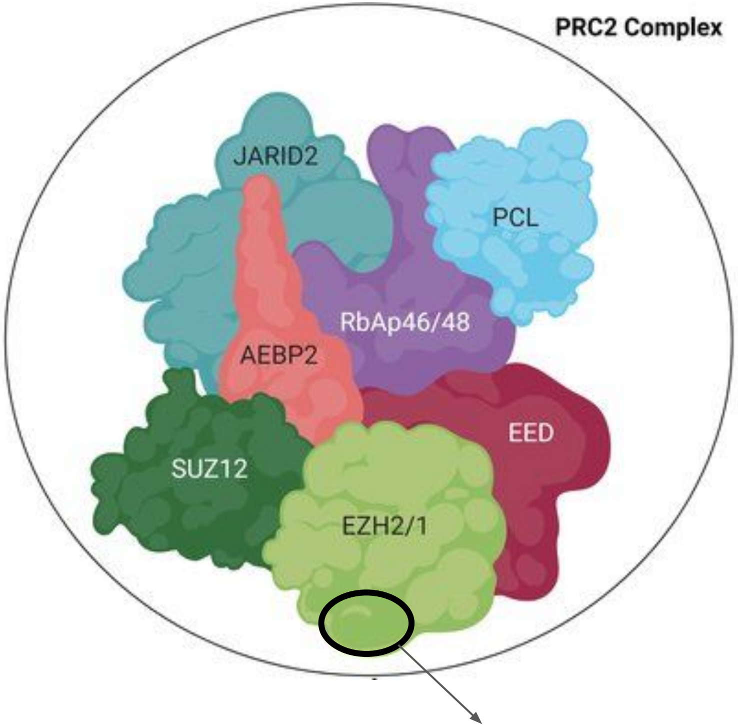
(https://xkcd.com/552)

Transcription-mediated histone modification



(Henikoff and Shilatifard 2011)

The example of H3K27me3, chiefly deposited by the polycomb repressive complex (PRC2)

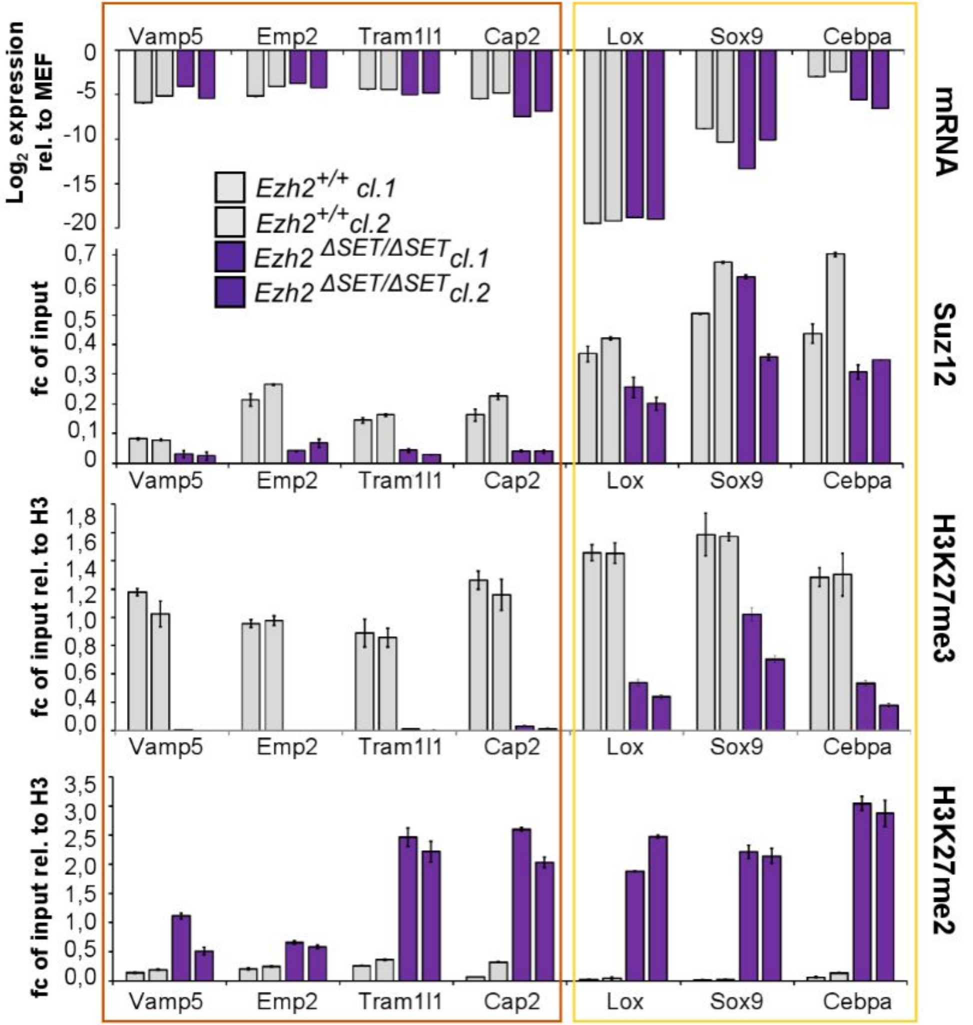


EHZ2’s SET domain catalyzes the addition of a 3rd methyl group to H3K27, i.e. H3K27me2 → H3K27me3

(Adapted from Zoroddu et al., 2021)

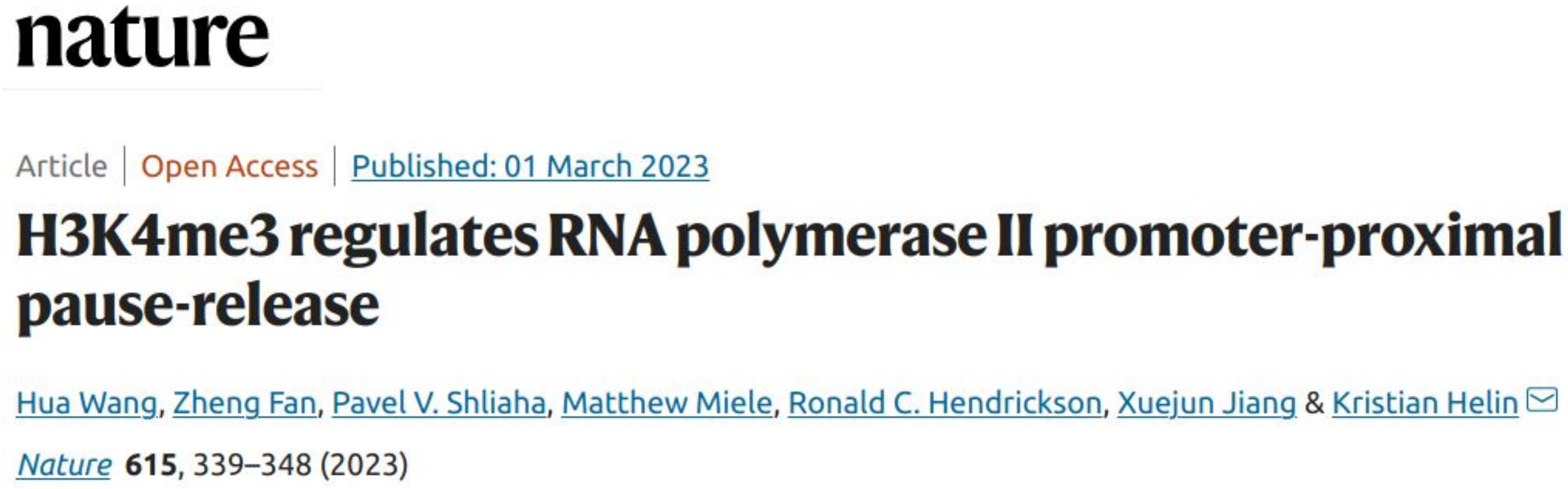
Abolishing the enzymatic activity of *Ezh2*, the gene responsible for depositing H3K27me3, abolishes (most of) the mark but does not prevent the repression of the target genes, nor cellular reprogramming

(Fragola et al., PLoS Genetics 2013)



Similarly, the loss of H3K4me3 appears to have no effect on nascent transcription

(Murray et al., bioRxiv 2019)



nature

Article | Open Access | Published: 01 March 2023

H3K4me3 regulates RNA polymerase II promoter-proximal pause-release

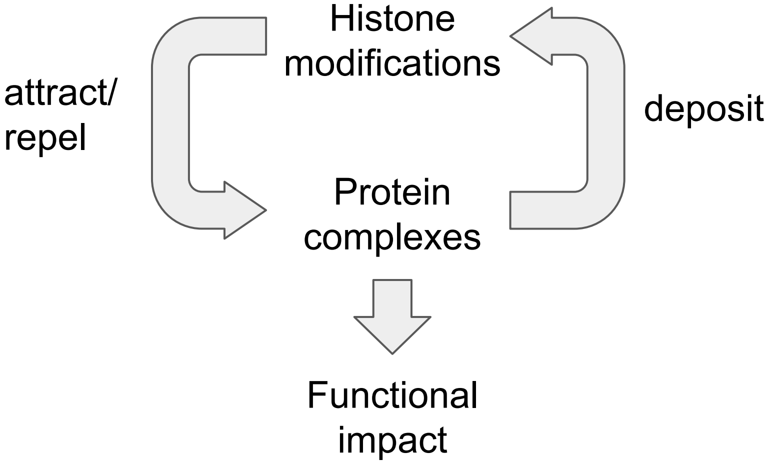
Hua Wang, Zheng Fan, Pavel V. Shliaha, Matthew Miele, Ronald C. Hendrickson, Xuejun Jiang & Kristian Helin

Nature **615**, 339–348 (2023)

“acute **loss of H3K4me3 does not have detectable effects on transcriptional initiation** but leads to a widespread decrease in transcriptional output, an increase in RNA polymerase II (RNAPII) pausing and slower elongation. We show that H3K4me3 is required for the recruitment of the integrator complex subunit 11 (INTS11), which is **essential for the eviction of paused RNAPII and transcriptional elongation**.”

Causality or correlation?

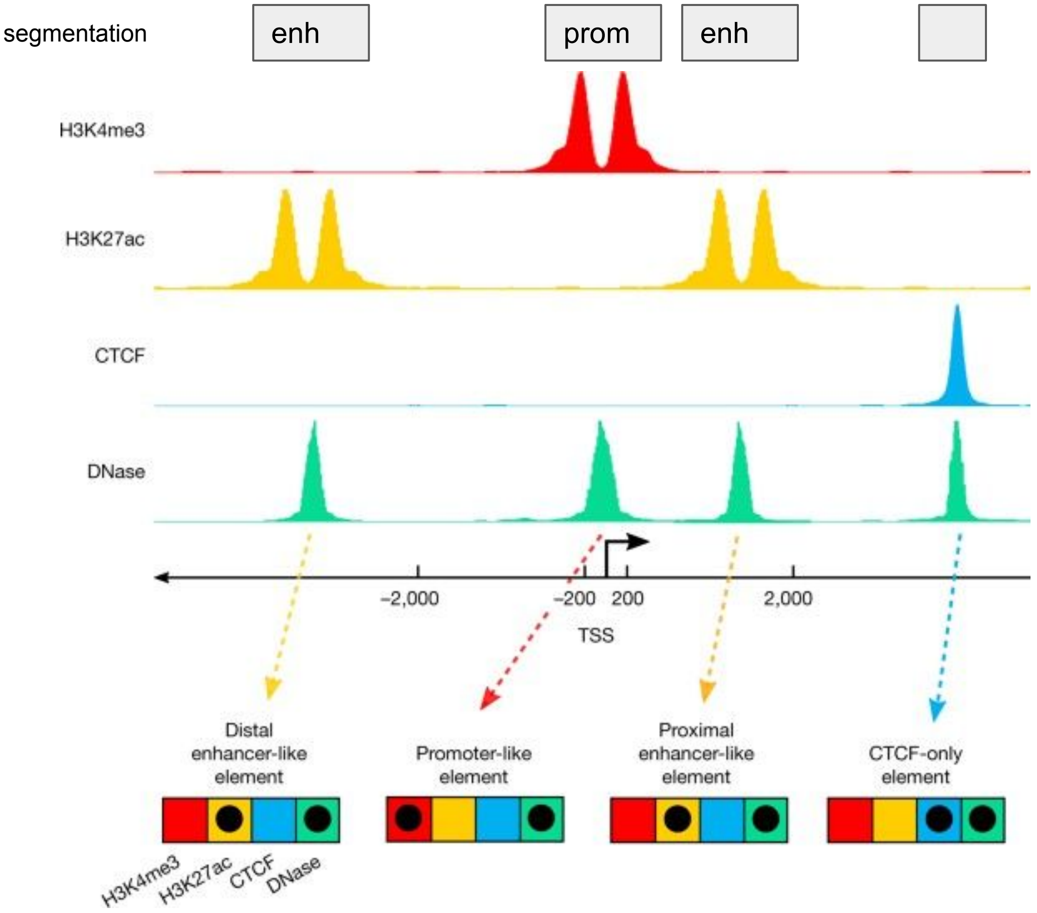
Most likely somewhere in the middle, depending on the modification/context



Whether they’re causative or not, they can serve as **proxies** for function.

This means that profiling a few histone modifications gives an overview of the epigenomic landscape of a cellular state which would otherwise require profiling all the potentially-relevant factors/complexes

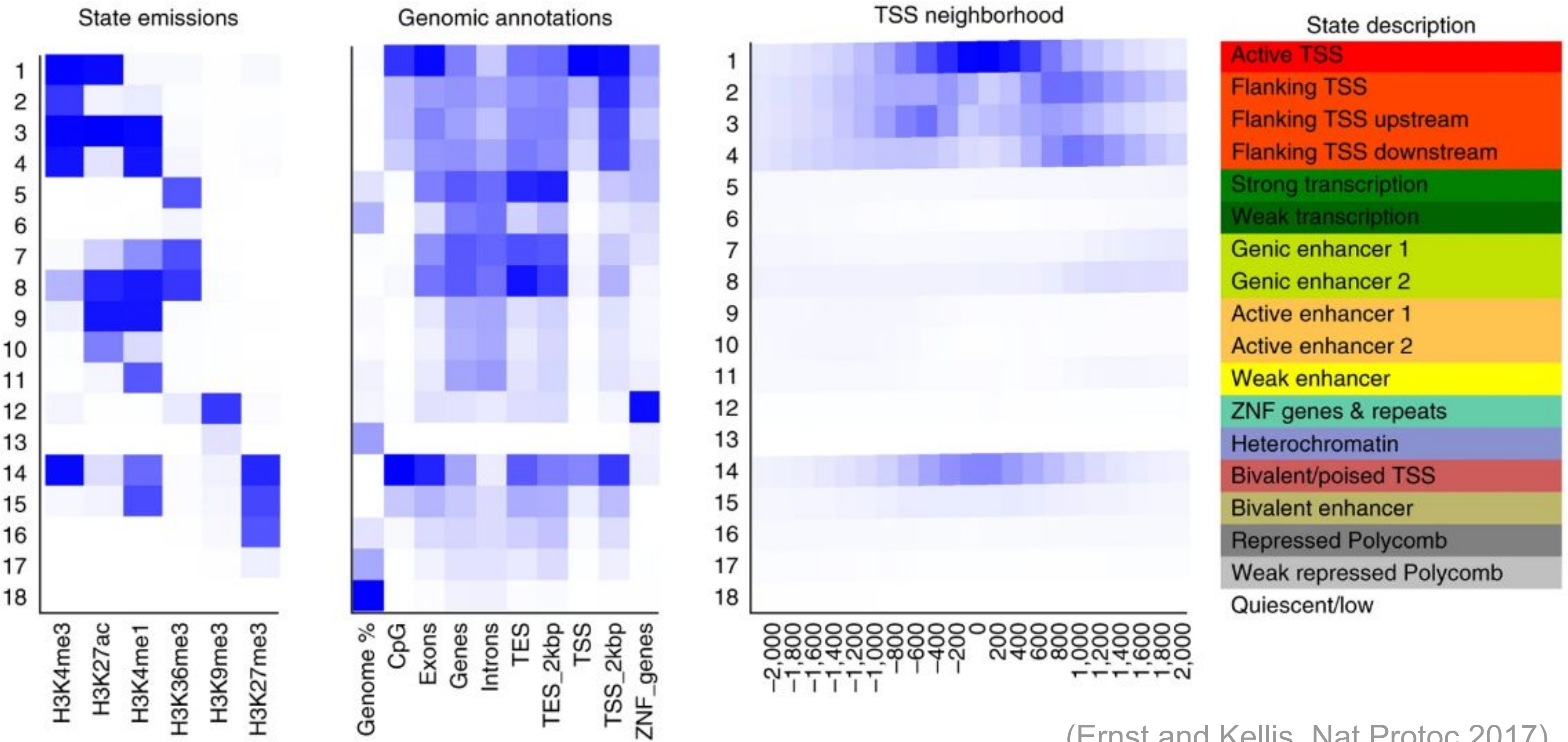
A signature-based encyclopedia of DNA elements



ENCODE's "signature strategy":

- Different types of functional genetic elements are associated with different chemical signatures
- We can identify functional elements by identifying these signatures genome-wide

So how many kinds of functional elements/states are there?



(Ernst and Kellis, Nat Protoc 2017)

Some stuff is pretty clear:

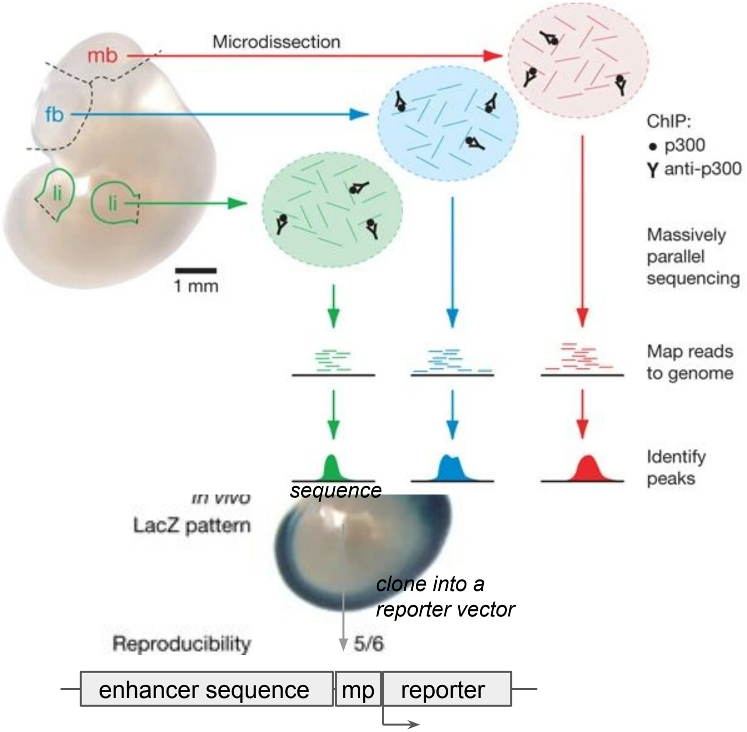
- **Transcription start site (TSS):**
  - **H3K4me3** is almost always associated with active/poised TSS
  - Active TSS are marked by **H3K27ac**
  - So-called “poised” (or bivalent) TSS are instead marked by both **H3K4me3** and **H3K27me3**

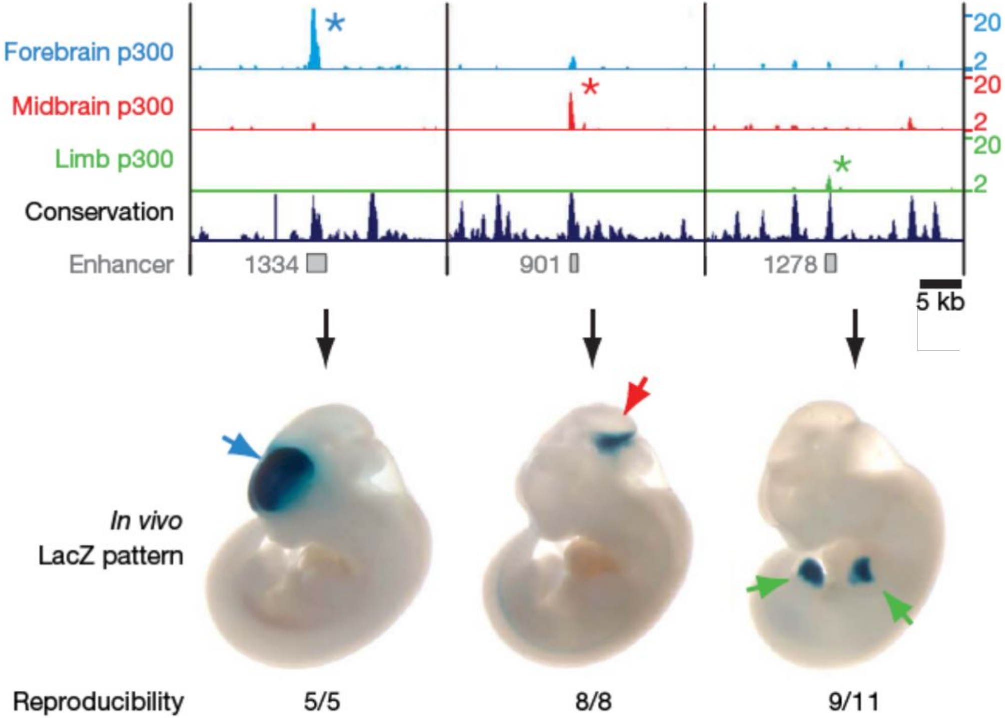
- **Enhancers:**
  - Most enhancers have **H3K4me1**
  - Active enhancers are marked by **H3K27ac**
  - So-called “poised” (or bivalent) enhancers are marked by **H3K4me1** and **H3K27me3**

- Repressed elements are marked by **H3K27me3**

- Heterochromatin is marked by **H3K9me3**
- Insulators: CTCF+cohesin

p300 and validation of enhancer activity

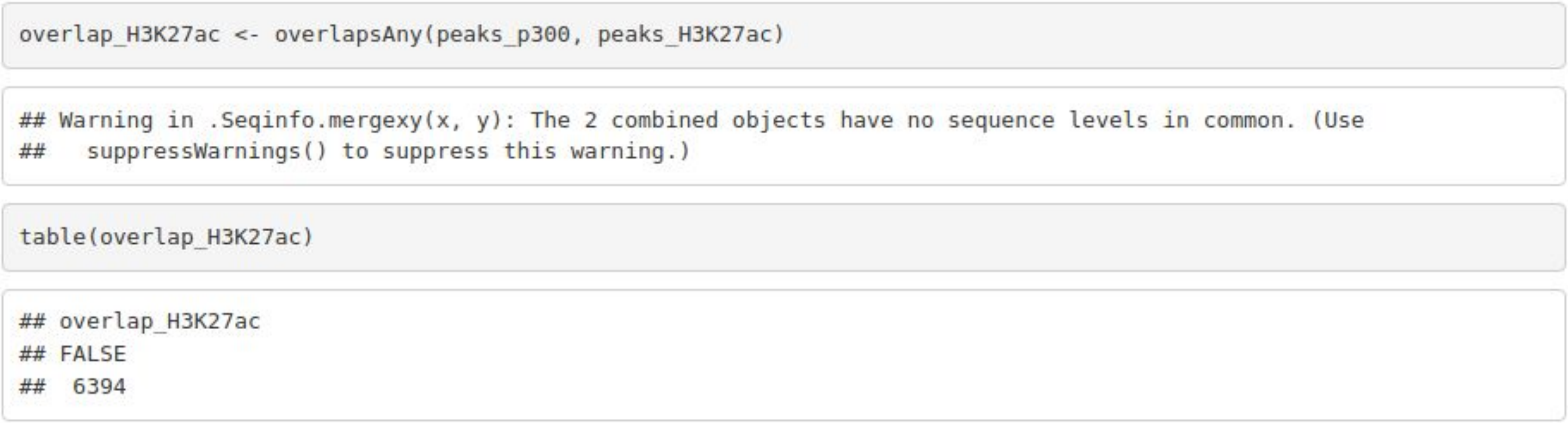




(Adapted from Visel et al., 2009)

vista enhancers: http://enhancer.lbl.gov

Inconsistent seqlevels



```
overlap_H3K27ac <- overlapsAny(peaks_p300, peaks_H3K27ac)
```

```
## Warning in .Seqinfo.mergexy(x, y): The 2 combined objects have no sequence levels in common. (Use
##   suppressWarnings() to suppress this warning.)
```

```
table(overlap_H3K27ac)
```

```
## overlap_H3K27ac
## FALSE
##  6394
```

Assignment

- Using the peaks you downloaded last week, identify bivalent domains (H3K27me3 + H3K4me3) in mouse embryonic stem cells (mESC)
    - Split those bivalent domains into those that overlap a TSS, and those that don’t.

- What happens to those regions upon differentiation?
    - Choose a differentiated cell type (e.g. hepatocytes, neural progenitor, or smooth muscle cells)
    - Download the H3K27me3 and H3K4me3 peaks from this cell type
    - How many of the mESC bivalent domains are, in this differentiated cell type, overlapping either mark or their combination?
        - Provide a separate answer for domains that overlap a TSS and those that don’t.